Решения

Задача 1. Дедлайн. 25 баллов

1. Оценки не изменились, в то время как их дисперсия уменьшилась в два раза. Это можно увидеть из матричного представления оценок МНК и их дисперсии. Действительно, консультант А. должен был получить $\hat{\beta} = (X'X)^{-1}X'y$, но вместо этого получил $\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y$ где $\tilde{X}$ это матрица размера $2n \times k$ (n — число наблюдений и k — число регрессоров)

$$\tilde{X} = \begin{bmatrix} X \\ X \end{bmatrix}$$

Заметим, что $(\tilde{X}'\tilde{X})^{-1} = (2X'X)^{-1} = \frac{1}{2}(X'X)^{-1}$ и $\tilde{X}'y = 2X'y$. Поэтому

$$\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y = \frac{1}{2}(X'X)^{-1}(2Xy) = \hat{\beta}$$

Далее, используя известный результат $\mathbb{V}ar(\hat{\beta}) = \sigma^2(X'X)^{-1}$, получим

$$\mathbb{V}ar(\tilde{\beta}) = \sigma^2(\tilde{X}'\tilde{X})^{-1} = \frac{1}{2}\sigma^2(X'X)^{-1} = \frac{1}{2}\mathbb{V}ar(\hat{\beta})$$

На практике же необходима несмещённая оценка ковариационной матрицы оценок коэффициентов:

$$\hat{\mathbb{V}ar}(\beta) = \hat{\sigma}^2(\tilde{X}'\tilde{X})^{-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k}(X'X)^{-1} = \frac{\sum_{i=1}^{124} e_i^2}{124-5}(X'X)^{-1}$$

Воспользуемся фактом, что в выборке Антона каждое наблюдение повторилось ровно 2 раза, то есть сумма квадратов остатков выросла в 2 раза:

$$\hat{\mathbb{V}ar}(\tilde{\beta}) = \hat{\tilde{\sigma}}^2(\tilde{X}'\tilde{X})^{-1} = \frac{\sum_{i=1}^{2n} e_i^2}{2n-k}(X'X)^{-1} = \frac{2\sum_{i=1}^{124} e_i^2}{248-5}(X'X)^{-1} = \frac{119}{243}\hat{\mathbb{V}ar}(\hat{\beta})$$

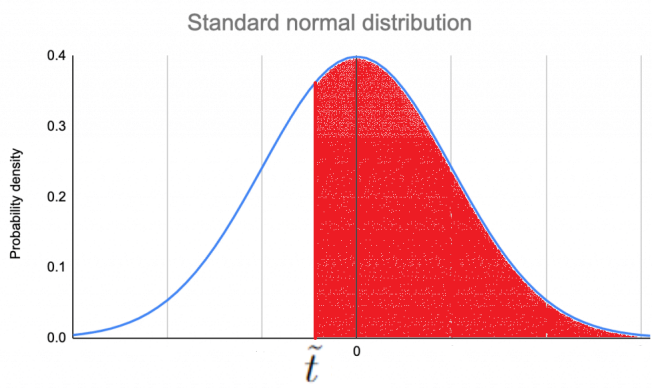
2. По расчётам Антона нулевая игпотеза не отвергается (судя по приведённому p-value). Консультанту Антону переделывать расчёты не стоит, так как ответ на вопрос, какую гипотезу принять, а какую отвергнуть – не поменяется. Пояснение: Для тестирования нулевой гипотезы должна была использоваться статистика

$$t = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{s.e.(\hat{\beta}_1 + \hat{\beta}_2)}$$

и нулевая гипотеза должна быть отвергнута если значение t превышает (положительное) критическое значение из таблицы. Поверялась односторонняя гипотеза! Однако вместо нее использовалась

$$\tilde{t} = \frac{\tilde{\beta}_1 + \tilde{\beta}_2 - 1}{s.e.(\tilde{\beta}_1 + \tilde{\beta}_2)} = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sqrt{\hat{\mathbb{V}ar}(\tilde{\beta}_1 + \tilde{\beta}_2)}} = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sqrt{0.489\hat{\mathbb{V}ar}(\hat{\beta}_1 + \hat{\beta}_2)}} = \sqrt{2.042} \cdot t$$

Предпоследнее равенство верно, потому что вся ковариационная матрица $\mathbb{V}ar(\hat{\beta})$ уменьшилась в два раза.



Далее, заметим что p-value равно 0.52 для данной гипотезы, значит расчётная статистика $\tilde{t}$, используемая Антоном, отрицательная. Схематично это изображено на графике. Тогда верная тестовая статистика $t = \frac{1}{\sqrt{2.042}} \cdot \tilde{t}$ тоже является отрицательной, значит вывод о том что нулевая гипотеза не отвергается, оказался верным.

Разбалловка:

Вопрос 1. 10 баллов

2 балла (1+1), если просто без обоснования написано, что оценки не изменились, а дисперсия уменьшилась в 2 раза.

8 баллов за обоснование: 4 за доказательство (например, как в приведённом решении) про оценки и 4 балла за вывод о дисперсии. Фактом про то, чему равна ковариационная матрица при стандартных предпосылках, можно пользоваться как готовым. Но если это тоже выводится в решении, то плюс 2 балла. Возможно получение баллов за верный ход доказательства, если не было перехода к несмещённой оценке дисперсии (корректировки на n-k). Возможны альтерантивные доказательства из геометрических соображений. Если этот пункт решался для парной регрессии (в заданиии множественная регрессия с 5 коэффициентами), то набранные баллы разделяются на 2.

Вопрос 2. 15 баллов

2 балла, если написан вывод Антона, что нулевая гипотеза не отвергается, судя по p-значению, равному 0.52.

2 балла, если просто дан ответ, что вывод Антона о гипотезах не поменяется.

3 балла за то, как выглядит расчётная t-статистика для проверки нулевой гипотезы. За F-статистику до 3 баллов в зависимости от подробности, хотя дальнейшие выкладки с ней затруднительны.

8 баллов за выкладки, что произошло с расчётной статистикой и как она соотносится с правильной расчётной статистикой.

Если ответ на этот вопрос дан не для односторонней, а для двусторонней гипотезы, то за весь этот пункт максимум 10 баллов из 15, если остальной ход решения верный.

Задача 2. Регрессия по двум точкам. 25 баллов

1. 7 баллов за первый пункт.

   Все предпосылки теоремы Гаусса-Маркова выполнены, значит $\hat{\beta}_{OLS}$ является несмещенной и эффективной. (2 балла. Но возможно, что человек и формально доказывал, тогда можно добавить баллов за верные выкладки) Разберемся с состоятельностью (5 баллов). В комментариях к условию в конце предлагалось пользоваться неравенством для докзательства состоятельности. Здесь приводится решение с использованием этого неравенства, но можно доказывать и иначе.

   (a) Возводя в квадрат обе части неравенства и используя классическое неравенство Маркова, получаем

   $$P(|X| > a) = P\left(X^2 > a^2\right) \leqslant \frac{\mathbb{E}\left(X^2\right)}{a^2}$$

   (b) Заметим, что $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Используя формулу для $\hat{\beta}_{OLS}$

   $$\begin{aligned}
   \hat{\beta}_{OLS} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \\
   &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y} \cdot \sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \\
   &= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} y_i \\
   &= \sum_{i=1}^{n} w_i y_i
   \end{aligned}$$

   где

   $$w_i = \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

   Тогда

   $$\sum_{i=1}^{n} w_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = 0$$

   и

   $$\sum_{i=1}^{n} w_i x_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i + \bar{x} - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = 1$$

   Наконец, подставляя $y_i = \alpha + \beta x_i + \varepsilon_i$ в полученное выражение для $\hat{\beta}_{OLS}$ и

используя свойства w_i, установленные выше, получим

$$\hat{\beta}_{OLS} = \sum_{i=1}^{n} w_i(\alpha + \beta x_i + \varepsilon_i) = \beta + \sum_{i=1}^{n} w_i \varepsilon_i$$

(c) Нужно показать что для любого $a > 0$

$$P(|\hat{\beta}_{OLS} - \beta| > a) \to 0$$

при $n \to \infty$. Согласно неравенству из пункта (a), необходимо понять, что происходит с величиной $\mathbb{E}\left((\hat{\beta} - \beta)^2\right)$. Раскрывая квардат суммы и используя $\mathbb{E}(\varepsilon_i \varepsilon_j) = 0$, получим

$$\mathbb{E}\left((\hat{\beta} - \beta)^2\right) = \sum_{i=1}^{n} w_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \to 0$$

исключая патологические случаи, когда начиная с какого-то j все x_j одинаковые. Таким образом, оценка $\hat{\beta}_{OLS}$ является состоятельной.

2. 8 баллов за этот пункт.

Обозначим взятую пару точек (x, y) и (x', y'). Прямая, проходящая через эти две точки, будет иметь наклон

$$\tilde{\beta} = \frac{y' - y}{x' - x}$$

Видим, что эта оценка является линейной по игреку (2 балла). Подставляя выражения для y' и y, получим

$$\tilde{\beta} = \frac{(\alpha + \beta x' + \varepsilon') - \alpha + \beta x + \varepsilon)}{x' - x} = \beta + \frac{\varepsilon' - \varepsilon}{x' - x}$$

Значит

$$\mathbb{E}(\tilde{\beta}) = \beta + \frac{1}{x' - x} \cdot \mathbb{E}(\varepsilon' - \varepsilon) = \beta$$

То есть $\tilde{\beta}$ является несмещенной оценкой β (2 балла). Так как все предпосылки теоремы Гаусса-Маркова выполнены, эффективной является оценка $\hat{\beta}_{OLS}$, а $\tilde{\beta}$ от нее отли- чается. Значит $\tilde{\beta}$ не является эффективной (формально нужно показать что ее дисперсия больше, но здесь это очевидно, она даже не зависит от n). (2 балла)

Конечно, $\tilde{\beta}$ не является состоятельной. Вероятность $P(|\tilde{\beta} - \beta| > a)$ не зависит от

n, не стремится к нулю и определяется распределением ε_i

$$P(|\tilde{\beta} - \beta| > a) = P\left(\frac{|\varepsilon' - \varepsilon|}{|x' - x|} > a\right) \neq 0$$

За (не)состоятельность тоже 2 балла.

3. 7 баллов за этот пункт.

   Так как каждая из $\tilde{\beta}_m$ является линейной и несмещенной, а $\tilde{\beta}$ представляет собой их линейную комбинацию с весами, складывающимися к единице, она также будет линейной и несмещенной (3 балла). Формально

$$\tilde{\beta} = \frac{1}{M}\sum_{m=1}^{M}\frac{y'_m - y_m}{x'_m - x_m}$$

   и

$$\mathbb{E}(\tilde{\beta}) = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}(\tilde{\beta}_m) = \frac{1}{M}\sum_{m=1}^{M}\beta = \beta$$

   При фиксированной выборке единственный источник неопределенности — какие наблюдения будут выбраны для построения $\tilde{\beta}$. Другими словами, при фиксированной выборке, рассмотрим случайную величину

$$Z_m = \frac{\varepsilon'_m - \varepsilon_m}{x'_m - x_m}$$

   которая принимает одно из $\frac{n(n-1)}{2}$ возможных значений

$$\left\{\frac{\varepsilon_1 - \varepsilon_2}{x_1 - x_2},\ \frac{\varepsilon_1 - \varepsilon_3}{x_1 - x_3},\ \ldots,\ \frac{\varepsilon_{n-1} - \varepsilon_n}{x_{n-1} - x_n}\right\}$$

   равновероятно. Заметим что

$$\tilde{\beta} = \frac{1}{M}\sum_{m=1}^{M} Z_m$$

   тогда по Закону Больших Чисел

$$\tilde{\beta} \ \xrightarrow{p}\ \mathbb{E}(Z_m) = \frac{1}{\frac{n(n-1)}{2}}\sum_{k=1}^{n(n-1)/2}\frac{\varepsilon'_k - \varepsilon_k}{x'_k - x_k}$$

   За верно найденный предел по вероятности 5 баллов.

4. 3 балла. Так как обе оценки $\hat{\beta}_{OLS}$ и $\tilde{\beta}$ являются линейными по игрек и несмещенными, и все предпосылки теоремы Гаусса-Маркова выполнены, можно сделать

вывод что $\hat{\beta}_{OLS}$ лучше, в том смысле что она обладает меньшей дисперсией.

Задача 3. Комбинация прогнозов. 25 баллов

1. 10 баллов. Ищем оптимальные веса:

$$e(\omega) = \omega \cdot e_1 + (1 - \omega) \cdot e_2$$

Так как по условию $E[e_1] = E[e_2] = 0$, то $E[e(\omega)] = 0$. Тогда:

$$MSE(\omega) = E[e(\omega)^2] = \omega^2 \cdot \sigma_1^2 + (1 - \omega)^2 \cdot \sigma_2^2 + 2\omega \cdot (1 - \omega) \cdot \sigma_{12} =$$

$\omega^2 \cdot (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) + \omega \cdot (2\sigma_{12} - 2\sigma_2^2) + \sigma_2^2$

Оптимальные веса находим из условий первого порядка (или как вершину параболы):

$$\omega^* = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

$$1 - \omega^* = \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

Проверка достаточного условия минимума: (2 балла)

$$(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) \geq (\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2) = (\sigma_1 - \sigma_2)^2 \geq 0$$

2. 10 баллов (4+6):

а) равные веса: (4 балла)

$$MSE(\omega = 0.5) = \omega^2 \cdot \hat{\sigma_1^2} + (1 - \omega)^2 \cdot \hat{\sigma_2^2} + 2\omega \cdot (1 - \omega) \cdot \hat{\sigma_{12}} =$$

$0.25 \cdot 6662 + 0.25 \cdot 6597 + 2 \cdot 0.25 \cdot 6618 = 6623{,}75$

б) Бейтс-Грейнджер: (3 + 3)

$$\omega^* = \frac{6597 - 6618}{6662 + 6597 - 2 \cdot 6618} = -0.91$$

$$1 - \omega^* = 1.91$$

$$MSE(\omega^*) = (-0.91)^2 \cdot 6662 + 1.91^2 \cdot 6597 + 2 \cdot (-0.91) \cdot 1.91 \cdot 6618 = 6577.83$$

3. 5 баллов.

Для сравнения качества прогнозов сравниваем MSE. (1 балл)

Комбинация прогнозов с весами по методу Бейтса-Гренджера является более предпочтительной, чем отдельный прогноз. Комбинация с равными весами лучше первого прогноза. (1 балл)

Лучшую комбинацию прогнозов обеспечивает тип весов по методу Бейтса-Гренджера (1 балл)

Тип весов по методу Бейтса-Гренджера всегда будет не хуже, чем веса по 0.5 (1 балл)

Оба способа будут приводить к одинаковым результатам тогда, когда дисперсии прогнозов будут равны друг другу. (1 балл)

Типичные ошибки:

В условии задания не было ограничения на $1 \geq \omega \geq 0$. Добавляя это ограничение, участники решали иную задачу, которая упрощала и изменяла ответы на следующие пункты. - минус 2 балла в первом пункте (если ошибка появлялась уже там), минус 1 балл за веса во втором пункте и -3 балла за MSE2 во втором пункте.

Если ошибка MSE получалось отрицательной, и это оставалось финальным ответом, то дополнительно -1 балл

Арифметические ошибки:
1) Если возникала арифметическая ошибка при подставлении в формулу, и если она не приводила к существенным искажениям последующих ответов, то оценка снижалась на 2 балла.
2) Если арифметическая ошибка приводила к изменению формулы оптимальных весов в первом пункте, но эта формула всё ещё оставалась логичной и не обязана была своим внешним видом насторожить участников, то за такую ошибку возникало снижение баллов от 4 до 7 (в зависимости от этапа её совершения).
3) Если формула выходила нелогичной (например, в формуле оптимального веса получилось, что вес первого прогноза положительно зависит от дисперсии первого прогноза), то минус 7 баллов из 10, а баллы за дальнейшие пункты делились на 2.

Задача 4. Переобучение безработных (25 баллов)

1. (10 баллов) Найдите оценку влияния переобучения безработных на заработную плату после трудоустройства способом, который предлагает Виктор.

   Сначала сопоставим пары. Виктор с помощью пробит-модели бинарного выбора оценил вероятность безработного фактически пройти программу переобучения в зависимости от пола и возраста: $Pr(D = 1) = (\beta_0 + \beta_1 * X_1 + \beta_2 * X_2)$, откуда нашёл коэффициенты для пробит-регрессии $\hat{\beta}_0 = -1.25$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1.2$

   В Таблице 1 указаны эти расчётные вероятности для проходивших перобучение. Это 0.5 для G1 и 0.84 для G2. Во таблице 2 расчётных вероятностей нет, зато есть оценки $\hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2$. И можно сопоставить пары не по расчётным вероятностям, а по значениям внутренней функции $\hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2$. Рассчитаем такие оценки для таблицы 1: Для G1 это -1.25+0.05*25+1.2*0=0, что совпадает с оценкой для группы A2 из таблицы 2. Для G2 это -1.25+0.05*21+1.2*1=1, что совпадает с оценкой для группы A3 из таблицы 2. Иными словами, для A2 и A3 расчётные вероятности тоже равны 0.5 и 0.84 соответственно.

   Поскольку в методе Виктора пары берутся 1 к 1, то необходимо взять из A2 все 30 человек (столько же в группе G1) и из A3 20 человек (столько же в группе G2). Поучаем 20+30 = 50 пар.

   Тогда расчёт оценки по методу Виктора (с округлением):

$$\hat{\alpha}_1 = \frac{1}{50} \cdot [20 \cdot (\frac{10}{0.5} - \frac{8}{1-0.5}) + 30 \cdot (\frac{9}{0.84} - \frac{2}{1-0.84})] =$$

   $[20 \cdot (20 - 16) + 30 \cdot (10{,}714 - 12.5)] \frac{1}{50} = (20 * 4 - 30 * 1.786)/50 = 0.52$

   Из 10 баллов: 4 балла за верное сопоставление групп из 1 и 2 таблиц, 6 баллов за оценку эффекта. Если размер групп неверный или не учтён, то 2 балла из 6 за оценку эффекта, если допущена арифметическая ошибка, то снимается 1 балл из 6 за оценку эффекта. Если сопоставления нет, а приведён некоторый расчёт по всем наблюдениям из двух таблиц, то не более 1 балла за весь этот пункт.

2. (4 балла. 1 за ответ, 3 за аргументы) Можно ли сказать, что выводы из оценки Виктора можно применить ко всем безработным в стране A? Почему?

   Нет. Судя по данным, приведённым в таблице 1, это молодые люди в возрасте 21-25 лет, и именно к ним подбирались «пары» из таблицы 2 по методу, который предложил Виктор. То есть оценка Виктора справедлива только для молодых

безработных, но не для остальных возрастов. Для них оценка может быть совсем иной.

3. (4 балла) Можно ли вместо способа Виктора посчитать оценку как разницу средних заработных плат между теми, кто прошёл и кто не прошёл переобучение? Почему?

   Нет, нельзя. Так оценивать некоректно хотя бы по причине «самоотбора»: программу могли проходить люди, которые отличаются по исходным характеристикам от тех, кто не проходил. Это же можно рассматривать и как проблему пропуска существенных переменных (т.к. сравнение средних по 2 группам даст тот же результат, что и парная регрессия Y от D), что влечёт за собой смещение оценки при D. Смещение оценки из простого сравнения средних отсутствовало бы, только если бы проводился «идеальный эксперимент», где безработных рандомизировали бы по 2 группам: кто проходит и кто не проходит обучение.

4. (7 баллов) Коллега Виктора сомневается в результатах его расчётов и считает, что между фактическим переобучением и будущей заработной платой есть эндогенность: существуют такие ненаблюдаемые характеристики человека, как мотивация и способности, а они могли повлиять и на решение переобучаться, и на будущую зарплату. Предложите способ оценки, который решал бы эту проблему. Какие предпосылки об имеющихся данных нужно сделать, чтобы воспользоваться этим методом?

   Проблему эндогенности можно решить методом инструментальных переменных (1 балл). В качестве инструмента нужен показатель, который бы одновременно коррелировал бы с фактическим прохождение переобучения D (свойство релевантности инструментальной переменной) и не коррелировал бы с ненаблюдаемыми характеристиками безработных (свойство эзогенности инструментальной переменной) (2 балла). В качестве такого показателя в данном исследовании можно использовать переменную Z, так как предложение от службы занятости с одной стороны, может коррелировать с фактом дальнейшего прохождения переобучения D и с другой стороны, не зависит от мотивации и способностей безработного, если сотрудники службы занятости при напрааленении на переобучение ориентируются только на объективные характеристики (2 балла). Итого оценка методом инструментальных переменных выглядит следующим образом (2 балла):

$$\hat{\alpha}_{IV} = \frac{\overline{Y}_{Z=1} = \overline{Y}_{Z=0}}{\overline{D}_{Z=1} = \overline{D}_{Z=0}}$$

   Можно перейти к терминологии потенциальных исходов и рассказать эту исто-

рию через оценку local average treatment effect (LATE), но в таком случае также необходимы подробные пояснения.